O‘ZBEKISTON MILLIY UNIVERSITETI XABARLARI, 2026, [15]
ISSN 2181-7324(Print)

FILOLOGIYA
http://journals.nuu.uz
Social sciences
ISSN 3093-8996(Onlayn)

UDK 81'33

Feruza Y. RAHMATOVA,
PhD, dotsent, Osiyo xalqaro universiteti, Buxoro, O‘zbekiston
E-mail: rahmatovaferuza516@gmail.com ORCID: 0009-0006-8002-1689

Bux DU professori, f.f.d G.Toirova taqrizi asosida

O‘ZBEK VA TURK TILLARI SOMATIK LEKSIKASINI KORPUS TADQIQOTIGA SARALASH VA KO‘P SATHLI LINGVISTIK ANNOTATSIYALASH MEZONLARI

Annotatsiya

Maqolada o‘zbek va turk tillari somatik leksikasini korpusda o‘rganishga tayyorlash uchun ma’lumot manbalarini tanlash, birliklarni saralash va ko‘p sathli lingvistik annotatsiyalash mezonlari tizimlashtiriladi. Dastlabki leksik inventar izohli lug‘atlar asosida shakllantirilgan: o‘zbek tilida 335 ta, turk tilida taxminan 230 ta somatik birlik qayd etilgan. Bu sonlar lug‘at materialidan tuzilgan nomzod birliklar ro‘yxatini bildiradi; ular matn korpusidagi chastota yoki ikki til bo‘yicha yakuniy qiyosiy ko‘rsatkich sifatida talqin qilinmaydi. Maqolada matn manbalarini tavsiflash, lemma-token-ma’no qatlamlarini farqlash, yozuv variantlarini normallashtirish, anatomik va ko‘chma ma’nolarni belgilash hamda annotatsiya ishonchliligini nazorat qilish tartibi taklif etiladi.

Kalit so‘zlar: somatik leksika, somatizm, korpus lingvistikasi, lingvistik annotatsiya, o‘zbek tili, turk tili, polisemiya, semantik tarmoq.

КРИТЕРИИ ОТБОРА СОМАТИЧЕСКОЙ ЛЕКСИКИ УЗБЕКСКОГО И ТУРЕЦКОГО ЯЗЫКОВ ДЛЯ КОРПУСНОГО ИССЛЕДОВАНИЯ И МНОГОУРОВНЕВОЙ ЛИНГВИСТИЧЕСКОЙ АННОТАЦИИ

Аннотация

В статье систематизируются критерии отбора источников, соматических единиц и многоуровневой лингвистической аннотации для корпусного исследования узбекской и турецкой лексики. Предварительный лексический инвентарь сформирован на основе толковых словарей: выделено 335 узбекских и около 230 турецких соматических единиц. Эти показатели отражают список словарных кандидатов, а не частотность в текстовых корпусах и не окончательные сопоставительные результаты. Предлагаются процедуры описания текстовых источников, разграничения леммы, токена и контекстного значения, нормализации графических вариантов, разметки прямых и переносных значений, а также контроля надежности аннотации.

Ключевые слова: соматическая лексика, соматизм, корпусная лингвистика, лингвистическая аннотация, узбекский язык, турецкий язык, полисемия, семантическая сеть.

CRITERIA FOR SELECTING THE SOMATIC VOCABULARY OF UZBEK AND TURKISH FOR CORPUS RESEARCH AND MULTILEVEL LINGUISTIC ANNOTATION

Annotation

This article systematizes criteria for selecting sources and somatic units and for applying multilevel linguistic annotation in corpus research on Uzbek and Turkish vocabulary. A preliminary lexical inventory was compiled from explanatory dictionaries: 335 Uzbek and approximately 230 Turkish somatic units were identified. These figures represent candidate lemmas from dictionaries; they are not corpus frequency counts or final cross-language comparison results. The article proposes procedures for documenting text sources, distinguishing lemmas, tokens and contextual senses, normalizing script variants, annotating literal and figurative meanings, and checking annotation reliability.

Keywords: somatic vocabulary, somatism, corpus linguistics, linguistic annotation, Uzbek language, Turkish language, polysemy, semantic network.

Kirish. Zamonaviy kompyuter lingvistikasi va tabiiy tilni qayta ishlash (Natural Language Processing – NLP) sohasida lingvistik resurslarni shakllantirish, ayniqsa muayyan leksik-semantik guruhlarni korpus ma’lumotlari asosida formallashtirish va modellashtirish dolzarb ilmiy vazifalardan biridir. Inson tana a’zolari va ular bilan bog‘liq tushunchalarni ifodalovchi somatik leksika (somatizmlar) barcha tillarda eng qadimiy, barqaror va yuqori funksional yuklamaga ega fundamental leksik qatlamni tashkil etadi. U inson tanasi, uning tashqi va ichki a’zolari, tana qismlari hamda ularning tuzilishiga daxldor nomlarni qamrab oladi.

Ushbu qatlamni kompyuter lingvistikasi nuqtai nazaridan tadqiq etishda lug‘atda qayd etilgan leksemalar ro‘yxatini tuzishning o‘zi yetarli emas. Leksemaning matndagi realizatsiyasi, grammatik shakli, leksik qurshovi va muayyan matnda faollashgan ma’nosi o‘rtasidagi aloqani qayd etish zarur. Shu sababli bu ishda korpus ma’lumoti oddiy matnlar yig‘indisi emas, balki manbasi ko‘rsatilgan, izlanishi mumkin bo‘lgan va tilshunoslik mezonlari asosida belgilashga tayyorlangan matnlar majmuasi sifatida tushuniladi.

Mavzuga oid adabiyotlar tahlili. Korpus resurslarini yaratish va lingvistik jihatdan belgilash masalalari o‘zbek kompyuter lingvistikasi doirasida alohida yo‘nalish sifatida yoritilgan. Elektron korpusni shakllantirishning lingvistik asoslari, korpus lingvistikasi va annotatsiyalash tamoyillariga oid ishlar ushbu maqoladagi manba tanlash hamda belgilash mezonlari uchun metodik tayanch bo‘lib xizmat qiladi [1; 4; 5]. Somatik birliklarning frazeologik va ma’noviy xususiyatlarini tavsiflovchi lug‘atlar esa nomzod birliklarni aniqlash va kontekstual ma’nolarni tekshirishda muhim [2; 3;].

Tadqiqot metodologiyasi. Korpus bazasini shakllantirish uchun leksikografik va matniy manbalarni tanlash mezonlarini belgilash;

Somatizmlarni anatomik, ontologik, diaxronik va iste’mol mezonlari bo‘yicha saralash;

Korpus matnlarini tozalash, normallashtirish va dublikatlardan tozalash qoidalarini belgilash;

Morfologik, sintaktik va semantik annotatsiya tamoyillarini hamda annotatsiya sifatini baholash tartibini asoslash.

Materiallar va metodlar

Manbalar

Korpus bazasini shakllantirish uchun taklif etilayotgan protokolda tilning funksional uslublari va tarixiy qatlamini qamrab olish tamoyiliga amal qilinadi. Maqola doirasida manbalar quyidagi guruhlarga ajratiladi:

Leksikografik va izohli lug‘atlar. O‘zbek tili bo‘yicha: 5 jildli «O‘zbek tilining izohli lug‘ati», «O‘zbek tili frazeologik lug‘ati», «O‘zbek tili o‘xshatishlarining izohli lug‘ati», «O‘zbek xalq shevalari lug‘ati» va «O‘zbek tili sinonimlarining katta izohli lug‘ati». Turk tili bo‘yicha: Türk Dil Kurumu (TDK) nashrlari – «Güncel Türkçe Sözlük», «Türkçede Benzetmeler Sözlüğü», «Türkçe Deyimler Sözlüğü», «Türkiye Ağızları Sözlüğü» va «TDK Büyük Türkçe Sözlük».

Korpus resurslari va matnlar massivi. O‘zbek tili Milliy korpusi va elektron matnlar bazalari [13]; turk tili korpus resurslari (TS Korpus, TNC – Türkçe Ulusal Derlemi). Turkcha resurslarning aniq manzil va versiya ma’lumotlari foydalanilgan platformaga ko‘ra bibliografiyaga kiritiladi.

Mumtoz va tarixiy manbalar. Mahmud Koshg‘ariyning «Devonu lug‘otit turk» asari matni.

Badiiy, publitsistik va ilmiy matnlar. O‘zbek va turk adabiyoti namoyandalari (A. Qodiriy, Cho‘lpon, Oybek, P. Qodirov, O‘. Hoshimov, T.Murod, X.E.Adivar, R.N.Guntekin, S.F.Abasiyanik va boshqalar) asarlaridan ajratilgan matn parchalari.

Somatik birliklarning dastlabki inventarini tuzishda izohli lug‘atlar asosiy leksikografik tayanch bo‘ldi: o‘zbek tili uchun «O‘zbek tilining izohli lug‘ati», turk tili uchun TDK «Güncel Türkçe Sözlük». Lug‘at nomzod lemmalar va ma’no tavsiflarini beradi, korpus esa ularning qo‘llanish dalillarini taqdim etadi. Lug‘at maqolasining mavjudligi birlikning real matniy chastotasini tasdiqlamaydi va lug‘atdagi ma’noni korpusdagi har bir matnga mexanik tatbiq etib bo‘lmaydi.

Tahlil va natijalar. Korpusga asoslangan yondashuvda uch daraja qat’iy farqlanadi: lug‘atdagi lemma, matndagi token va matnga bog‘langan ma’no. Masalan, bosh – lemma, boshim, boshida, boshlarni – so‘zshakllar (tokenlar); bosh og‘rig‘i va ishning boshi birikmalarida esa bir lemmaning turlicha semantik talqinlari namoyon bo‘ladi. Bu farq saqlanmasa, so‘zshakllar soni leksemalar soni bilan, leksema chastotasi esa muayyan ma’no chastotasi bilan aralashib ketishi mumkin.

Saralash mezonlari

Somatizmlarni to‘rtta asosiy mezon bo‘yicha tizimlashtirish taklif etiladi: (I) anatomik joylashuv, (II) ontologik xususiyat, (III) qo‘llanish davri (diaxronik mezon) va (IV) iste’mol darajasi. Dastlabki inventar uchun asosiy mezon birlikning inson tanasi, tana qismi yoki ichki a’zoni nomlash bilan bevosita semantik aloqadorligidir. «Somatik leksika» tushunchasining qamrovi tadqiqotlarda turlicha belgilanishi mumkinligi sababli markaziy somatizmlar, tuzilmaviy somatik nomlar va chegaraviy biologik birliklar (masalan, qon, suyak, teri, mushak) farqlanishi kerak.

Matnlarni tayyorlash

Taklif etilayotgan tayyorlash bosqichida matnlardan HTML belgilar, navigatsiya qismlari, reklama, takroriy sarlavhalar, sahifa raqamlari va lisoniy ma’lumotga tegishli bo‘lmagan elementlar ajratiladi. Tinish belgilari va defislar ommaviy o‘chirilmaydi, chunki ular leksik chegara, qo‘shma nom va sintaktik bog‘lanishni aniqlashda muhim. Har bir tahrir amali jurnalga yozilib, asl matnni tiklash imkoniyati saqlanishi kerak.

O‘zbek tilida lotin va kirill yozuvlari uchun umumiy qidiruv qatlamini yaratish, asl grafik shaklni esa alohida maydonda saqlash taklif etiladi. Transliteratsiyada o‘, g‘, q, h harflarining mosligi hamda tutuq belgisi va apostrofning turli Unicode variantlari tekshirilishi lozim (masalan, qo‘l, qo'l, qo‘l, qo’l shakllari bir lemmaga bog‘lanadi). Turk tilida ğ, ş, ç, ö, ü, ı va İ harflarining kodlash xususiyatlari hisobga olinadi; ı/i va I/İ juftliklarini umumiy kichik harfga o‘tkazish xato lemmalar hosil qilishi mumkinligi sababli har bir til uchun alohida normallashtirish moduli belgilanadi.

Dublikatlarni bartaraf etishda aynan takrorlangan hujjatlar va bir manbaning qisman qayta chop etilgan parchalari farqlanishi kerak. Hujjat darajasida xesh-identifikatorlardan, parcha darajasida matniy o‘xshashlik tekshiruvidan foydalanish; takror deb topilgan holatni chiqarishdan oldin uning manba aloqasini metama’lumotda qayd etish taklif qilinadi.

Annotatsiya sathlari va sifat nazorati

Taklif etilayotgan annotatsiya protokoli tokenlashtirish, lemmatizatsiya, morfologik, sintaktik va semantik sathlardan iborat bo‘ladi. Ishonchlilikni baholash uchun sinov to‘plami, ikki mustaqil annotator va kelishuv koeffitsiyentidan foydalanish tavsiya etiladi. Bu maqolada annotatorlararo kelishuvning amaliy qiymati hisoblanmagan; koeffitsiyent natijasi faqat real belgilash tajribasi o‘tkazilgach berilishi kerak.

Dastlabki leksik inventar

«O‘zbek tilining izohli lug‘ati» asosida 335 ta, «TDK Güncel Türkçe Sözlük» bo‘yicha esa taxminan 230 ta somatik leksema nomzodi ajratilgan. Bu miqdorlar lug‘at maqolalaridan shakllantirilgan inventar ko‘rsatkichlaridir. Matn korpuslari hajmi va token chastotalari maqolada hisoblanmaganligi sababli ularni real qo‘llanish chastotasi sifatida talqin qilish mumkin emas. Turk tiliga doir son taxminiy bo‘lgani uchun, yakuniy qo‘lyozmada lemmalarni kiritish mezonlari va ro‘yxat manbasi aniq ko‘rsatilishi lozim.

Anatomik joylashuv bo‘yicha taqsimot

Dastlabki o‘zbekcha inventar anatomik joylashuv bo‘yicha uch guruhga ajratildi. Jadvaldagi turkcha birliklar ma’no muvofiqligini keyinchalik tekshirish uchun keltirilgan namunalardir; ularning soni bo‘yicha qiyosiy hisob ushbu jadvalda berilmagan:

Guruh	Ulushi	Soni	Misollar (o‘zbekcha/turkcha)	Semantik xususiyati
Boshdagi a’zolar	43%	145	bosh/baş, ko‘z/göz, qosh/kaş, yuz/yüz, quloq/kulak, burun/burun, til/dil, og‘iz/ağız, tish/diş, jag‘/çene	Kognitiv va ko‘chma ma’nolar alohida tekshiriladi
Tashqi/tanadagi a’zolar	41%	136	qo‘l/el-kol, oyoq/ayak, bilak/bilek, tirsak/dirsek, tizza/diz, tovon/topuk, bo‘yin/boyun, yelka/omuz	Harakat va ijtimoiy faoliyat ma’nolari tekshiriladi
Ichki a’zolar	16%	54	yurak/kalp-yürek, igar/karaciğer, o‘pka/akciğer, oshqozon/mide, uyrak/böbrek, ichak/bağırsak	Emotsional ma’nolar bilan aloqasi tekshiriladi

Ontologik xususiyat bo‘yicha taqsimot.

Somatizmlarni ontologik jihatdan tavsiflashda anatomik denotatli birliklar hamda inson holati yoki tushunchasini ifodalovchi, somatiklik chegarasi alohida asoslanishi lozim bo‘lgan birliklar farqlanishi mumkin.

Ikkinchi guruhdagi birliklarni asosiy inventarga kiritish mezoni alohida izohlanishi kerak:

Diaxronik taqsimot.

Mahmud Koshg‘ariyning «Devonu lug‘otit turk» asari tarixiy qatlamni aniqlash uchun manba sifatida olinadi. Tarixiy birliklar zamonaviy shakl va ma’nolar bilan qiyoslanganda shakliy davomiylik, fonetik o‘zgarish, arxaiklashish va zamonaviy tilda qayd etilmaslik kabi tavsifiy toifalar qo‘llanadi. Quyidagi jadval ushbu toifalarni ko‘rsatadi; miqdoriy ulushlar, korpus chastotasi va aniq qiyosiy natijalar asar nashri, sahifalar hamda tekshirilgan birliklarning yakuniy ro‘yxati bilan tasdiqlangandan keyingina e’lon qilinishi lozim:

Tavsifiy toifa	Mezon	Misollar
Shakli va ma’nosi davom etgan	Tarixiy va zamonaviy shakl hamda ma’no yaqinligi	ko‘z/köz, til/til, tish/tiş, yuz/yüz
Fonetik o‘zgarishga uchragan	Tovush mosligi kuzatiladi	adaq – oyoq, ag‘iz – og‘iz, tamur – tomir
Arxaiklashgan	Tarixiy manbada bor, bugun cheklangan	alun, ernek, yag‘in, yulun
Zamonaviy adabiy tilda qayd etilmagan	Zamonaviy muqobil alohida tekshiriladi	chichamuq, urung qaraq, bo‘kseg

Iste’mol darajasi bo‘yicha taqsimot.

Iste’mol darajasini belgilash uchun korpusda birliklarning janrlar va davrlar bo‘yicha chastotasi hisoblanishi lozim. Bosh, ko‘z, qo‘l, oyoq, yurak kabi birliklar faol nomzodlar sifatida, bebilchak, solinchak, taqim, pochin, qayshar, o‘taka, mardum kabi birliklar esa cheklangan yoki tarixiy-dialektal nomzodlar sifatida tekshiriladi. Bu tasnif chastota hisoblari bilan tasdiqlanmaguncha dastlabki faraz hisoblanadi.

Tokenlashtirish. Matn so‘zshakllar, tinish belgilari va zarur hollarda ko‘pso‘zli birliklarga ajratiladi. Agglyutinativ xususiyat tufayli somatik lemmalar ko‘plab qo‘shimchali shakllarda uchraydi (ko‘zlarimizdan, gözlerimizden: ko‘plik, egalik va chiqish kelishigi ko‘rsatkichlari), biroq tashqi o‘xshashlik qo‘shimchalar vazifasining mutlaq bir xilligini bildirmaydi.

Lemmatizatsiya. Derivatsiya natijasida mustaqil leksemaga aylangan birliklar avtomatik ravishda shu lemmaning tokeni deb hisoblanmaydi.

Morfologik annotatsiya. So‘z turkumi, son, egalik, kelishik va zarur hollarda shaxs belgilari qayd etiladi; universal belgilash tizimlari har bir tilning grammatik o‘ziga xosligiga moslashtiriladi. Morfologik annotatsiya semantik teglashga tayyorlovchi bosqich bo‘lib, uning o‘rnini bosmaydi (masalan, boshida joylashuvni ham, vaqtning boshlanishini ham anglatishi mumkin).

Sintaktik annotatsiya. Birlikning bosh so‘zga tobeligi, aniqlovchi yoki to‘ldiruvchi vazifasi, fe’l bilan birikuvi va barqaror konstruktsiyadagi o‘rni qayd etiladi. Sintaktik belgi semantik ma’noning o‘zi emas, balki qaror uchun dalillardan biri hisoblanadi. Tillararo qiyosda sintaktik qoliplar mexanik ko‘chirilmaydi.

Kontekst oynasi. Dastlab maqsadli birlik joylashgan gap olinadi; ma’no oldingi yoki keyingi gapga bog‘liq bo‘lsa, kengaytirilgan diskurs matni saqlanadi.

Semantik annotatsiya. Har bir maqsadli token uchun lemma, til, matn, lug‘atdagi ma’no nomzodi va annotator tanlagan ma’no identifikatori o‘zaro bog‘lanadi. Birlamchi tasnif anatomik bevosita ma’no, metaforik ko‘chish, metonimik ko‘chish, barqaror birikma tarkibidagi yaxlit ma’no va yetarli dalil bo‘lmagan holatni farqlaydi. Masalan, «bosh – rahbar» munosabati metaforik talqin qilinishi mumkin, «qo‘l – barmoq» esa butun–qism (meronimik) aloqadir va ma’no ko‘chishi bilan bir darajada emas.

Qaror qabul qilish tartibi. Avval lemma va so‘z turkumi tekshiriladi, so‘ng anatomik ma’noning mosligi baholanadi, keyin lug‘atdagi boshqa ma’no variantlari bilan solishtiriladi, zarur bo‘lsa frazeologik lug‘at yoki kengroq diskursga murojaat qilinadi. Hech bir ma’no ishonchli tanlanmasa, token majburan teglanmaydi, unga «noaniq» belgisi va izoh kiritiladi.

Xulosa. O‘zbek va turk tillari somatik leksikasini korpus sharoitida o‘rganish uchun taklif etilgan mezonlar lug‘aviy inventarni manbali matn dalillari bilan bog‘lashga xizmat qiladi. Dastlabki lug‘at tanlanmasida o‘zbek tilidan 335 ta, turk tilidan taxminan 230 ta somatik birlik nomzodi ajratilgan; bu qiymatlar korpus chastotasini anglatmaydi. Somatizmlarni anatomik joylashuv, ontologik maqom, tarixiy qatlam va qo‘llanish xususiyatlari bo‘yicha tavsiflash mumkin. Taklif etilgan ko‘p sathli annotatsiya sxemasi lemma, token, morfologik shakl, kontekstual ma’no va tillararo muvofiqlikni alohida maydonlarda qayd etadi.

ADABIYOTLAR

1. Abdurahmonova N.Z. O‘zbek tilining elektron korpusini yaratishning lingvistik asoslari: Filol. fan. dokt. (DSc) diss. – Toshkent, 2018. – 280 b.
2. Eshonqulov B. O‘zbek tilida somatik frazeologizmlarning semantik-pragmatik xususiyatlari: Filol. fan. nomz. diss. – Samarqand, 2004. – 156 b.
3. Rahmatullaev Sh. O‘zbek tilining izohli frazeologik lug‘ati. – Toshkent: O‘qituvchi, 1978. – 380 b.
4. Rahimov S. Kompyuter lingvistikasi va lingvistik annotatsiyalash asoslari. – Toshkent: Akademnashr, 2021. – 210 b.
5. Hamroev M. Korpus lingvistikasi: nazariya va amaliyot. – Toshkent: Fan, 2020. – 195 b.
6. O‘zbek tilining izohli lug‘ati. 5 jildli. J. 1–5. – Toshkent: «O‘zbekiston milliy ensiklopediyasi» Davlat ilmiy nashriyoti, 2006–2008.
7. Rahmatullaev Sh. O‘zbek tilining frazeologik lug‘ati. – Toshkent: Qomuslar bosh tahririyati, 1992. – 320 b.
8. Mahmudov N., Xudoybeganova D. O‘zbek tili o‘xshatishlarining izohli lug‘ati. – Toshkent, 2013.
9. O‘zbek xalq shevalari lug‘ati. – Toshkent: Fan, 1971. – 410 b.
10. Azizov O. va boshq. O‘zbek tili sinonimlarining katta izohli lug‘ati. – Toshkent: O‘qituvchi, 2015. – 512 b.
11 Güncel Türkçe Sözlük. – Ankara: Türk Dil Kurumu Yayınları, 2021.– 1250s.
12 Büyük Türkçe Sözlük. – Ankara: Türk Dil Kurumu Yayınları, 2019. – 1450s.
13 O‘zbek tili Milliy korpusi. Elektron resurs. URL: https://uzcorp.uz (murojaat sanasi: 12.10.2026).
14 Choriev A. Badiiy matnlarda somatizmlarning uslubiy vazifalari // O‘zbek tili va adabiyoti. – Toshkent, 2019. – № 4. – B. 110–118.